DOCUMENT RESUME

ED 086 726 TM 003 377

AUTHOR	Jensen, Arthur R.
TITLE	An Examination of Culture Bias in the Wonderlic Personnel Test.
PUB DATE	73
NOTE	25p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Caucasians; *Cross Cultural Studies; *Intelligence Tests; Item Analysis; Negroes; *Racial Differences; *Test Bias; *Test Validity
IDENTIFIERS	*Wonderlic Personnel Test

ABSTRACT

Internal evidence of cultural bias, in terms of various types of item analysis, was sought in the Wonderlic Personnel Test results in large, representative samples of whites and Negroes totalling some 1,500 subjects. Essentially, the lack of any appreciable Race X Items interaction and the high interracial similarity in rank order of item difficulties lead to the conclusion that the Wonderlic shows very little or no evidence of cultural bias with respect to the present samples, which, however, differ appreciably in mean scores. The items which best measure the g factor within each racial group are, by and large, the same items that show the largest interracial discrimination. (Author)

FILMED FROM BEST AVAILABLE COPY

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 086726

TM 003 272

An Examination of Culture Bias in the Wonderlic Personnel Test

Arthur R. Jensen

University of California, Berkeley

ABSTRACT

Internal evidence of cultural bias, in terms of various types of item analysis, was sought in the Wonderlic Personnel Test results in large, representative samples of whites and Negroes totalling some 1,500 subjects. Essentially, the lack of any appreciable Race × Items interaction and the high interracial similarity in rank order of item difficulties lead to the conclusion that the Wonderlic shows very little or no evidence of cultural bias with respect to the present samples, which, however, differ appreciably in mean scores. The items which best measure the g factor within each racial group are, by and large, the same items that show the largest interracial discrimination.

An Examination of Culture Bias in the Wonderlic Personnel Test

Arthur R. Jensen[1]
University of California, Berkeley

Psychometricians are generally agreed that a population difference in average test score is not, by itself, evidence of biased sampling of test items such as to favor (or disfavor) a particular cultural group. The mean difference between groups may be explainable in terms of factors other than culture bias in the item content of the test. Evidence of culture bias thus depends upon criteria other than a group mean difference.

There are two main classes of criteria for assessing test bias: external and internal. They are complementary. The external criteria are the more important in terms of the practical usefulness of the test and where predictive validity for a specific quantifiable performance criterion is possible. Bias is indicated when two (or more) populations show significantly different regressions of criterion measures on test scores. If the regression lines for the two (or more) groups do not differ significantly in intercept and slope, the test can be said to be "fair" to all groups with respect to the given criterion of external validity. Refinements and variations of this general external criterion for assessing test bias have been discussed extensively in the measurement literature (e.g., Cleary, 1968; Darlington, 1971; Humphreys, 1973; Jensen, 1968; Linn, 1973; Thorndike, 1971).

Internal criteria of cultural bias become important when discussing the construct validity of the test and in assessing claims of bias even when the external validity criteria give no evidence of bias. Such claims of test bias are sometimes made on the grounds that the external criterion of the test's validity is itself culture-biased and is therefore predictable by a culture-biased test. Internal criteria of bias get around this argument by examining the degree to which different socioeconomic and cultural groups differ in terms of various "internal" features of the test involving item statistics. The main criterion for the detection of bias lies in the magnitude of the groups × items interaction relative to other sources of variance in an analysis of variance (ANOVA) design comprised of Groups (G), Items (I), Subjects within Groups (S), and the interactions G × I and S × I. This method was first used by Cleary and Hilton (1968), who examined the G × I interaction on two forms of the Preliminary Scholastic Aptidude Test in white and Negro groups. The Race × Items interaction proved statistically significant but contributed to minimally relative to the main effects that the authors concluded: " . . . given the stated definition of bias, the PSAT for practical purposes is not biased for the groups studied." Stanley (1969) later showed that a considerable amount of the Race × Items interaction was due to just a few items that were too difficult in both racial groups and therefore did not discriminate much between them. Negroes scored rather uniformly lower than whites on most of the items.

The Groups × Items interaction is analyzable into two effects: (a) the similarity in the rank order of the percent passing, p, each item in each of the groups, and (b) the similarity between the groups in the differences between the p values of adjacent items in the test, i.e., p_1-p_2, p_2-p_3, etc. There are here called p decrements. Group differences in rank order

of item difficulties are termed disordinal interactions. Group differences in p decrements, when the rank order of p values is the same in both groups, are termed ordinal interactions. A measure of similarity between groups, such as the Pearson correlation between the groups, in p values and p decrements, can serve as sensitive indexes of the degree to which the groups behave differently with respect to different items. Presumably all test items in any test are not equally culture biased, and to the degree that items differ in this property, the extent of cultural differences between two groups relevant to performance on the test should be related inversely to the size of the intergroup correlations of p values and of p decrements. Also, if more test items are culturally irrelevant or unreliable in one group than in another, this can be expected to result in different magnitudes of the test's internal consistency reliability in the two groups.

The present study examines the Wonderlic Personnel Test (WPT) for evidence of culture bias in terms of these internal criteria when applied to representative white and Negro samples. The WPT is an obviously culture-loaded test of general intelligence. The fact that it is culture-loaded only means that most of the items are based on specific information and cognitive skills that are commonly acquired in present-day English-speaking western culture. This is obvious simply from inspection of the test items. Whether the obvious culture loading of the items biases the test to the disadvantage of any particular population with respect to another population is a separate question which can be answered only in terms of empirical investigation of test data from the groups in question.

The cultural-educational loading of the Wonderlic would seem to make it suspect as a possibly culture-biased test in the American Negro population. This should be a point of concern when the WPT is used in business and

industry, and especially where precise external criteria of the WPT's validity in the white and Negro groups is not available. More than 6,500 organizations routinely use the WPT as a part of their personnel selection and placement procedures, making it one of the most widely used tests of mental ability.

Detailed descriptions of the WPT and references to previous research can be found in Buros (1972, pp. 724-6). Briefly, the WPT is a group-administered paper-and-pencil test of 50 verbal, numerical, and spatial items arranged in spiral ominbus fashion. It is generally given with a 12-minute lime limit. Alternate form reliabilities average .95. Use of the WPT is claimed to have validity where educability or trainability is a job requirement (Wonderlic & Wonderlic, 1972, p. 60). Large representative samples of males and females show no significant difference in total raw score on the WPT.

Negro Norms

Norms based on 38,452 Negro job applicants have been published (Wonderlic & Wonderlic, 1972). The authors state: "The vast amount of data studied in this report confirms that a very stable differential in raw scores achieved by Negro applicant populations exists. Where education, sex, age, region of country and/or position applied for are held constant, Negro-Caucasian WPT score differentials are consistently observed. These mean score differentials are . . . about one standard deviation apart when comparisons of Caucasians and Negroes are studied" (p. 3). As the authors note (p. 68), the Negro (as well as white) norms are based on biased samples of the Negro (and white) populations to the extent that they are based on an applicant population of individuals who are looking for jobs. The age group from 20 to 24 is predominantly represented for both sexes and for both races.

The published norms show the mean and median test score of Negro and white applicants for each of 80 different occupational categories, from the professional-managerial level to unskilled labor. The correlation between the Negro and white medians across the 80 occupational categories is .84 (the correlation between means is .87), indicating a high degree of similarity between the racial groups in their self-selection for various occupations. In other words, the rank order of median and mean test scores of applicants for various jobs is very similar in the Negro and white populations, despite the approximately 1 σ race difference in mean scores for all job categories.

Is there internal evidence in the test data that the 1 σ difference between whites and Negroes is attributable in whole or in part to culture bias in the WPT?

Method

Subjects

Parallel analyses were performed on two pairs of white and Negro samples. Thus the findings from the main analyses are replicated in two sets of Negro-white comparisons based on samples selected in different ways.[2]

Sample 1 consists of 544 white and 544 Negro Ss representing a random sample of the nationwide population of job applicants on which the published white and Negro norms are based for Form IV of the WPT. These large samples thus closely approximate the score distributions of the normative white and Negro populations, which have been given full statistical description in the manual of norms of the WPT (Wonderlic & Wonderlic, 1972). The samples were drawn without selection for characteristics such as age, education, job category, sex, and region. All Ss coded as "other minority" or Ss with Spanish surnames were excluded from the sample. In terms of the white σ

(standard deviation), the mean scores of the white and Negro samples differ by 1.05 σ as compared with 1.00 σ in the total normative populations.

Sample 2 consists of randomly slected test protocols of 204 white and 204 Negro Ss who were job applicants for entry level positions in a single company in New York City. No selection was made on age, education, and sex. Ss coded as "other minorities" and Spanish surnames are not included in the white sample. The white and Negro means of Sample 2 are very close to the national norms, but the SDs are almost double. (Sample 2: White $\bar{x}$ = 22.07, SD = 14.86; Negro $\bar{x}$ = 15.63, SD = 13.89. National Norms: White $\bar{x}$ = 23.32, SD = 7.50; Negro $\bar{x}$ = 15.80, SD = 7.06). In terms of the white sample SD, therefore, the Sample 2 white-Negro mean difference is only 0.43 σ, although it is 0.86 σ in terms of the normative white.

Results

P Values and P Decrements

The p value is the proportion of the total sample who answer a given test item correctly. P values were obtained for items 1 - 50 in the white and Negro groups.

The p decrement is the difference between the p values of ordinally adjacent test items, e.g., p_1-p_2, p_2-p_3, etc., where the subscript indicates the item number in the test. P decrements between adjacent items 1-2, 2-3, . . .,49-50 were obtained in both samples.

Table 1 shows the mean p values within sets of 10 items (and for all items) for each of the racial groups in Samples 1 and 2. The item p values

- - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - -

Table 1

Summary of Wonderlic Item Statistics on Sample 1 (N = 544 Whites and 544 Negroes) and Sample 2 (N = 204 Whites and 204 Negroes)

Items	Mean P				Correlation (Uncorrected) Between White P × Negro P		Correlation (Corrected for Attenuation) Between White P × Negro P		Correlation (Uncorrected) Between W × N P Decrements		Correlation (Corrected for Attenuation) Between W × N P Decrements	
	Sample 1		Sample 2									
	White	Negro	White	Negro	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
1-10	.815	.623	.829	.653	.886	.920	.901	.935	.907	.934	.923	.950
11-20	.662	.409	.682	.485	.802	.702	.837	.733	.845	.768	.884	.803
21-30	.461	.233	.439	.266	.879	.945	.895	.962	.855	.928	.876	.952
31-40	.238	.101	.230	.143	.937	.943	.963	1.032	.673	.789	.780	.914
41-50	.035	.007	.031	.014	.765	.933	.835	1.0192	.765	.938	.827	1.014
All Items	.442	.275	.442	.312	.932	.956	.939	.962	.792	.832	.820	.862

were correlated between racial groups within 10-item sets and over all 50 items. As can be seen in Table 1, these correlations are quite high even within sets of 10 items. This means that the relative difficulty of the items, as indicated by the proportion passing, is highly similar in the white and Negro samples.

The reliability of the p values within each racial group was estimated by obtaining the correlations between the p values of the same racial groups in Samples 1 and 2. These within-race correlations between p values are all over .90 and for all 50 items the correlations (or reliability of the p values) are .995 for whites and .992 for Negroes. Using the reliabilities thus obtained, the interracial correlations between item p values were corrected for attenuation, as shown in Table 1. The fact that the correlations after correction for attenuation are distributed about a mean of less than 1.00, of course, indicates that the interracial correlation of p values is significantly less than the intraracial correlation. Yet the corrected interracial correlations are very high, which means that the relative item difficulties, though not identical, are much alike in the white and Negro groups.

The p decrements were treated in exactly the same way. Since p decrements, unlike p values, are not systematically correlated with the item's ordinal position in the test, the interracial correlation between p decrements is a more sensitive index of group similarity than the correlation of p values. A high interracial correlation between p decrements means that the relative differences in difficulty between adjacent items are much alike in the two racial groups. If some items were more racially-culturally biased than others, resulting in different relative difficulties for whites and Negroes, it would be reflected in a low interracial correlation

between item p decrements, both with or without correction for attenuation. As can be seen in Table 1, this is not the case. The interracial correlations of p decrements are remarkably high. They are distributed about a mean of less than 1.00, however, which means that there is a slight but significant difference in the relative p decrements of the white and Negro groups.

P Values and P Decrements for Attempted Items Only

As the WPT is a timed test, very few Ss attempt every item. The typical pattern of response for most Ss is to answer the first 10 or 15 items and then to begin to skip around looking for items that appear relatively easy for them in order to obtain the highest score they possibly can in the time available. Items which were left unanswered by the S are considered to be not attempted.

Table 2 shows (a) the mean proportion of each group attempting items (in sets of 10 items), (b) the interracial correlation (corrected for attenuation) between these proportions, (c) the mean proportion, P_A, passing the attempted items (d) the interracial correlations of P_A, and (e) the correlation between proportion attempting and proportion passing the items.

- - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - -

Whites and Negroes are highly similar in the proportions attempting each item. The similarity is even greater for the proportion of each group

Table 2

Proportion Passing, P_A, of All Ss Who Attempted Items, and the Inter-racial Correlations

Items	Mean Proportion Attempting Items				White × Negro Correlation[1] Proportion Attempting		Mean Proportion Passing Attempted Items, P_A				Correlation[2] Between White P_A × Negro P_A		Correlation[1] Between White P_A × Negro P_A		Correlation Between % Attempting and % Passing Item			
	Sample 1		Sample 2		Sample	Sample	Sample 1		Sample 2		Sample	Sample	Sample	Sample	Sample 1		Sample 2	
	White	Negro	White	Negro	1	2	White	Negro	White	Negro	1	2	1	2	White	Negro	White	Negro
1 - 10	1.000	1.000	.935	.886	--[3]	.687	.815	.623	.881	.726	.886	.882	1.074	1.069	--[3]	--[3]	.202	.378
11 - 20	.995	.985	.823	.858	.981	.970	.665	.415	.744	.568	.797	.261	1.100	.360	.077	.047	.513	-.143
21 - 30	.906	.805	.789	.679	.989	.980	.434	.287	.545	.393	.882	.973	1.073	1.183	.043	.140	.087	-.025
31 - 40	.578	.383	.444	.364	.996	.987	.395	.232	.501	.375	.877	.866	1.052	1.039	.407	.615	.255	.244
41 - 50	.161	.088	.095	.051	.994	.961	.217	.081	.291	.242	.805	.643	.981	.783	.284	.079	.346	.019
All Items	.728	.653	.617	.568	.901	.824	.509	.328	.592	.461	.916	.796	1.162	1.009	.724	.711	.716	.578

[1] Corrected for attenuation

[2] Uncorrected.

[3] Since 100% of both racial groups attempted items 1-10, the computation of r is precluded.

passing the attempted items; the interracial correlations, when corrected for attenuation, generally do not differ significantly from unity. Overall, in both Samples 1 and 2, the interracial correlations of item difficulties of attempted items is so high as to indicate that the items have essentially the same relative difficulties in the white and Negro groups.

White-Negro Differences According to Type of Items

It is often claimed that Negroes perform relatively less well on verbal items than on other types, since presumably verbal content allows wider scope for cultural variations and the effects of bias on Negro scores. To see if this notion holds true for the various kinds of item content in the WPT, items were classified as shown in Table 3 and the mean White-Negro difference in these item categories was determined.

- - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - -

Since items in different categories occur unsystematically at different ordinal positions in the test and have different overall levels of difficulty in both racial groups, it was necessary, in order to make the appropriate comparisons, to transform the proportion passing to an index of item difficulty which constitutes an interval scale. As explained by Guilford (1954, pp. 418-419), this is accomplished by expressing the proportion passing in terms of the z score deviations of the normal curve. The group mean difference is thus expressed in σ of z score deviations. For example, if on a given item Group A has 84% passing and Group B has 60% passing, the

Table 3

Classification of Wonderlic Items and Mean Z Scale Difference Between White and Negro Groups[1]

Item Category	Number of Items	White-Negro Difference: All Items Sample 1 Mean	SD	All Items Sample 2 Mean	SD	Attempted Items Sample 1 Mean	SD	Attempted Items Sample 2 Mean	SD
Verbal Ability									
Opposite-Similar (single words)	14	.656	.25	.587	.21	.639	.24	.545	.28
Oddity Problem (single words)	3	.395	.39	.690	.12	.650	.39	.653	.12
Proverbs	5	.450	.14	.366	.12	.316	.26	.264	.40
Scrambled Sentences	2	.735	.21	.365	.19	.655	.29	.335	.13
Sentence Meaning	3	.500	.17	.343	.15	.536	.20	.303	.17
Total Verbal	27	.606	.25	.514	.21	.561	.27	.463	.29
Numerical Reasoning									
Number Series	3	.837	.03	.583	.15	.870	.11	1.270	.84
Arithmetic Reasoning (Money)	5	.694	.39	.350	.34	1.000	.36	.808	1.07
Arithmetic Reasoning (Quantity & Rate)	5	.658	.33	.402	.26	.576	.43	.020	.93
Total Numerical	13	.713	.30	.425	.27	.807	.38	611	1.03
Logical Reasoning									
Verbal Syllogisms	6	.452	.27	.308	.14	.367	.19	-.130	.89
Spatial-Geometric Reasoning	2	.850	.04	.295	.42	1.270	.65	.120	.41
Total Logical Reasoning	8	.551	.29	.305	.19	.593	.51	-.069	.78
Factual Information	1	.000	-	.240	-	.000	-	.000	-
Clerical Accuracy	1	.520	-	.340	-	.340	-	.210	-
All Items	50	.611	.28	.445	.23	.615	.36	.402	.67

[1] Sample 1: N = 544 whites and 544 Negroes
Sample 2: N = 204 whites and 204 Negroes

corresponding z scores (from the table of areas under the normal curve) are +1.00 and +0.25 and the difference between Group A and Group B is 1.00 - .25 = .75 σ By thus transforming p values to z scores, items of different difficulty in the two groups can be compared on an interval scale, permitting direct comparisons of the mean White-Negro z score differences for different types of items.

Table 3 shows the mean z scale difference between the white and Negro group on the various types of items, as well as the SD of the N items of each type. Because of the small numbers of items in the separate categories, the most important comparisons are between the totals for Verbal, Numerical, and Logical Reasoning. Also, more weight probably should be given to the results for attempted items. In Sample 1 there were no individual items with negative z values, either for all items or for attempted items, and there were only five such items among those attempted in Sample 2; in all cases these were items attempted by fewer than 8% of either group. That is to say, whites did better on all items attempted by more than 8% of Ss in either group. There is no regular tendency for the White-Negro difference to be greater for the verbal than for numerical or logical reasoning, and the smallest differences are in factual information and the interpretation of proverbs, which, surprisingly, are the types of items that are so often held up as examples of culture-loaded test items. There is no consistent difference between "all items" and "attempted items." Overall the White-Negro difference is about as great for the attempted items as for all the items. The rather low degree of consistency between results for Samples 1 and 2 would seem to make unwarranted any strong conclusions from the analysis in Table 3. What it does illustrate is the lack of any marked or consistent tendency for any one type of item to be more

racially discriminating than other types, as the items are here classified.

If specific type of content is not systematically related to the item's racial discriminability, is there any item characteristic that is so related? It was hypothesized that items' g loadings (or loading on the first principal component) when the item intercorrelation matrix is factor analyzed within each racial group separately would be most higly related to the item's discriminability between the racial groups. That is to say, the more highly an item is correlated with the general factor common to all items, within either racial group, the more highly it will discriminate between the racial groups. To test this hypothesis, the items' loadings on the first principal component (the g factor of the item intercorrelation matrix) were obtained from separate principal components analyses of the white and Negro data (Sample 2). The items' factor loadings were correlated with the items' z index of interracial discriminability (Table 3), for all items, not just attempted items. The Pearson correlation is .47 in the White sample and .62 in the Negro sample. For items with g loadings of greater than .40, the mean White-Negro z difference is .64 (for factor loadings in White sample) and .67 (for factor loadings in Negro sample); while for items with g loadings of less than .40, the corresponding z differences are .36 and .37, respectively. A similar relationship holds also for attempted items. The White-Negro z difference for all items with loadings of more than .40 on g is .52 (in White sample) and .66 (in the Negro); the corresponding figures for items loaded less than .40 are .35 and .31. When this was cross-validated in Sample 1, the White-Negro z difference for all items with g loadings greater than .40 is .78 (for White sample) and .79 (for Negro sample); the z differences for all items with g loadings less than .40 are .54 (White sample) and .55 (Negro sample). The cross-validating correlation between the Sample 2 factor loadings and the Sample 1

White-Negro *z* differences are .44 (White sample) and .33 (Negro sample). What all this means is that there is a substantial relationship between the size of the item loadings on the general factor common to all items in the Wonderlic and the magnitude of the White-Negro difference on the item, and this is true whether the *g* factor is determined in the White or in the Negro Sample. Neither the loadings on any components other than the first principal component (i.e., *g*) nor type of item content reveals any systematic relationship to the item's interracial discriminability. On the other hand, the items that best measure the general factor *within* each racial group are the same items, by and large, that discriminate most highly *between* the racial groups.

Analysis of Variance: Items × Subjects Matrix

The Race × Items interaction in a complete ANOVA of the Items × Subjects matrix provides a sensitive index of item bias relative to other sources of variance. Using the Sample 2 data, three such ANOVAs were performed: (1) on the total white and Negro groups, (2) on white and Negro groups equated on total WPT score, and (3) on "pseudo-racial" groups comprised entirely of two groups of white *S*s selected so that their total WPT score distributions closely match the normative white and Negro distributions in means and *SD*s. The ANOVAs for each of these conditions are summarized in Table 4. To that the three analyses can be directly compared,

- - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - -

the sum of squares for each source in the ANOVA is converted to omega

squared (ω^2) × 100, which is the percent of the total variance attributable to the given source.

For the ANOVA of the total white and Negro samples, all of the effects are significant beyond the .001 level, including the Race · Items interaction. But once the statistical significance of this interaction is shown, more important than statistical significance is the magnitude of the interaction relative to other sources of variance. The smaller it is, the more "fair" the test as regards culture bias. The appropriate index of "fairness," thus defined, is the A/B ratio, which, in terms of ω^2 is A = R/S and B = (R×I)/(I×S). In terms of F, A/B = $F_R/F_{R\times I}$. The two formulas for the A/B ratio are algebraically equivalent. If the Race · Items interaction is non-significant, it is presumed that no bias has been demonstrated and there is no point in computing the A/B ratio. The lower the value of the A/B ratio, the easier it would be to equalize or reverse the racial group means by item selection. Obviously a small group mean difference along with a large Groups × Items interaction would mean that a somewhat different selection of items from the same item population could equalize or reverse the group means. The higher the value of A/B, the less is the possibility of equalizing the group means through item selection from a similar population of items. This would not rule out the possibility of introducing different kinds of items into the test, but if doing so decreases the A/B ratio (even though it decreases the group mean difference), it can be argued that the minimizing of the group mean difference is simply a result of balancing item biases. Some tests equate male and female scores on this basis, balancing items that favor one sex with the selection of items that favor the other. Such a test, resulting in little or no mean sex difference but a large Sex × Items interaction, of course precludes the

use of such a test for studying the question of sex differences in the ability which the test purports to measure. The same thing would be true of any test which was made to equalize racial group differences at the expense of greatly increasing the Race × Items interaction. The desirable condition is to minimize the interaction as much as possible.

The A/B ratio for the total samples (Table 4) is 10.84. For comparison, a similar study of white and Negro elementary pupils showed an A/B ratio of 7.10 on the culture-loaded Peabody Picture Vocabulary Test and of 17.32 on the culture-reduced Raven's Progressive Matrices (Jensen, in press).

ANOVA on Equated White and Negro Samples. In a previous study, it was found that when groups of white and Negro school children were roughly matched for mental age (rather than chronological age), and ANOVA of the Peabody Picture Vocabulary Test (PPVT) items was performed, the Race × Items interaction was greatly reduced from its magnitude when the two racial groups were of the same chronological age but different mental ages (Jensen, in press). This finding suggests that a large part of the Race × Items interaction is attributable to a mental maturity × items interaction rather than to a racial-cultural difference per se. And this hypothesis was strengthened by showing that the same magnitude of the actual Race × Items interaction could be achieved entirely with the white sample, simply by dividing it into two "pseudo-racial" groups for the ANOVA. One group of white Ss was selected so that their distribution of total PPVT scores matched the Negro distribution in mean and SD; the other group of white Ss was selected so that its PPVT score distribution matched the total white distribution. When these two culturally homogeneous groups, corresponding to the Negro and white samples, were subjected to the same ANOVA as was

applied to the true racial groups, it reproduced the same results almost perfectly, including the Race × Items interaction. In other words, an interaction of this magnitude could be attributed to an average ability difference between the groups rather than to a cultural difference.

The same kind of analysis is here applied to the Wonderlic data. Since mental age is not a meaningful scale in an adult population, Negro and white Ss were simply matched for total score on the WPT. Perfect matching was possible on 127 White-Negro pairs, making the white and Negro total score distributions identical.

If the WPT items are culture-biased for Negroes, one might expect that whites and Negroes with the same total scores would obtain them in different ways, so that even when the main effect of Race is zero in the ANOVA, the Race × Items interaction would remain.

Table 4 shows the results of the ANOVA on the equated samples. The main effect of race was, of course, forced to be zero by equating the groups.

- - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - -

But note that the Race × Items interaction is very small and nonsignificant ($F = 1.25$, $df = 48/12,096$, $p > .10$). This finding is consistent with the hypothesis that the R × I interaction in the ANOVA of the total samples is due to the average difference in ability between the groups rather than to a cultural difference. It seems less likely that equating the white and Negro groups for total score should wipe out an R × I interaction if it truly reflected a cultural difference between the white and Negro groups.

Table 4

Omega Squared ($\omega^2 \times 100$) and F from ANOVA of Wonderlic Test in Total and Equated White and Negro Samples, and in "Pseudo-Race" Samples

	Total Samples			Equated Samples			"Pseudo-Race" Samples		
Source of Variance	df	$\omega^2 \times 100$	F	df	$\omega^2 \times 100$	F	df	$\omega^2 \times 100$	F
Race (R)	1	1.83	84.87	1	0	0	1	2.08	58.37
Items (I)	48	34.22	256.54	48	37.95	172.86	48	39.36	150.57
Subjects within Race (Ss)	406	8.75	7.75	252	6.45	5.56	194	6.87	6.51
R × I	48	1.04	7.83	48	0.29	1.26[a]	48	0.94	3.61
Ss × I	19,488	54.17		12,096	55.31		9,312	50.73	

[a] Nonsignificant, $p > .10$.

One might argue that white and Negro Ss who attain the same total socre must be highly similar in cultural background and therefore would show no significant R × I interaction. But are they culturally more similar than individuals of the same racial group who differ by 7 points in total Wonderlic score? (The σ of total scores in the normative white population is close to 7.) Siblings reared together in the same family differ by almost as much. Since the white and Negro population means differ by close to 1 σ (or 7 points on the WPT), we can do an ANOVA on a "pseudo-race" comparison by making up two groups of white Ss selected so that their score distributions closely approximate those of Negroes and whites. This was done by ranking all white scores from highest to lowest, and then, working in from both ends of the distribution, selecting pairs of Ss who differ by exactly 7 points in total score. The means of the two distributions differ by 7 points and they have the same SD = 12.78.

Table 4 shows the ANOVA of these "pseudo-race" groups. It can be seen that the results resemble the true racial comparison (Table 4--Total Samples), especially as regards the R × I interaction, which for the Total Samples constitutes 1.04% of the variance and for the "pseudo-racial" samples is 0.94%. The F for the R × I interaction is significant beyond the .001 level for both the Total Sample and Pseudo-Race Sample, and the A/B ratios are 10.84 and 16.31, respectively. The ratio of ω^2 for the interactions (R × I / Ss × I) is .019 in both the Total Sample and the "Pseudo-Race" Sample. All this indicates that a large part of the R × I interaction can be attributed to a level-of-ability × items interaction, since it is shown to exist in the "pseudo-race" groups which are both comprised of white Ss differing in average ability. If the significant R × I interaction were explainable only in terms of cultural differences

between the white and Negro groups, it seems highly improbably that it could be reduced to nonsignificance simply by equating the racial groups for overall level of ability, or that the same significant interaction could be produced within a culturally homogeneous white sample divided into high and low ability groups with overlapping score distributions similar to the total white and Negro distributions. In brief, from these three ANOVAs shown in Table 4, it would be extremely difficult to make a case that the Race × Items interaction is attributable to cultural bias. These analyses should have produced markedly different results if the popular claims of culture bias were in fact valid.

Discussion and Conclusion

Several different analyses of test item characteristics have failed to reveal evidence of culture bias for large Negro and white samples on the Wonderlic Personnel Test. If some items were more culture biased than others with respect to the cultural backgrounds of Negroes and whites, one should expect (a) significantly different rank order of p values (percent passing) for various items in the white and Negro samples, (b)significantly different intervals (i.e., p decrements) between the p values of adjacent test items in white and Negro samples, (c) a significant Race × Items interaction in the analysis of variance of the Race × Items × Subjects score matrix, even when both racial groups are equated for total score, and (d) systematic differences in the types of item content that discriminate most and least between the white and Negro samples. None of these expectations was borne out by the present data. The small but significant Race × Items interaction could be reduced to nonsignificance by equating the white and

Negro groups for overall score, which would not be expected if the two groups differed culturally in reaction to the test items. Moreover, it was possible to produce a significant "Pseudo-Race" × Items interaction within the culturally homogeneous white group simply by dividing the total white sample into two groups, one which duplicates the mean and SD of the Negro norms and the other which duplicates the mean and SD of the white norms. This suggests that the Race × Items interaction is really an ability level × items interaction rather than an interaction due to cultural differences.

The only way one could view these findings as being not incompatible with the hypothesis that the Wonderlic is a culturally biased test for Negroes would be to claim that culture bias depresses Negroes' performance on all the test items to much the same degree, which seems highly unlikely for cultural effects per se, and especially considering the great variety of item content in the Wonderlic. Otherwise it should be possible to make up subscales consisting of items on which the Negro group on the average does as well or better than the white group. This, however, is not possible with the present pool of Wonderlic items. The items that best measure the general factor common to all items within each racial group are also the same items that discriminate the most between the racial groups.

The present analyses yield no consistent or strong evidence that the Wonderlic is reacted to in any way differently in the Negro and white samples, except in overall level of performance, in which the normative populations differ by about one standard deviation. The present evidence lends no support to the hypothesis that the cause of this difference in average score on the Wonderlic is explainable in terms of cultural bias.

References

Buros, O. I. (Ed.) Seventh Mental Measurements Yearbook. Vol I. Highland Park, New Jersey: Gryphon Press, 1972.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cleary, T. A., & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Darlington, R. B. Another look at "cultural fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Guilford, J. P. Psychometric Methods. 2nd ed. New York: McGraw-Hill, 1954.

Humphreys, L. G. Implications of group differences for test interpretation. Assessment in a Pluralistic Society. Proceedings of the 1972 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1973. Pp. 56-71.

Jensen, A. R. Another look at culture-fair tests. In Western Regional Conference on Testing Problems, Proceedings for 1968, "Measurement for Educational Planning." Berkeley, Calif.: Educational Testing Service, Western Office, 1968. Pp. 50-104.

Jensen, A. R. How biased are culture-loaded tests? Genetic Psychology Monographs, in press.

Linn, R. L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-161.

Stanley, J. C. Plotting ANOVA interactions for ease of visual interpretation. Educational and Psychological Measurement, 1969, 29, 791-797.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Wonderlic, E. F., & Wonderlic, C. F. Wonderlic Personnel Test: Negro Norms. Northfield, Illinois: E. F. Wonderlic & Associates, Inc. 1972.